# Functional characteristics of novel pancreatic Pax6 regulatory elements

# Functional characteristics of novel pancreatic Pax6 regulatory elements

Adam Buckle[1], Ryu-suke Nozawa[1], Dirk A. Kleinjan[2*], Nick Gilbert[1*]

[1]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Crewe Rd, Edinburgh, EH4 2XR, UK

[2]Centre for Mammalian Synthetic Biology, University of Edinburgh, Kings buildings, Edinburgh, EH9 3FF, UK

Correspondence:

DA Kleinjan

E-mail: dirk-jan.kleinjan@ed.ac.uk

N Gilbert

Tel: +441316518551; Fax: +441316518800; E-mail: nick.gilbert@ed.ac.uk

## Abstract

Complex diseases, such as diabetes, are influenced by comprehensive transcriptional networks. Genome-wide association studies have revealed that variants located in regulatory elements for pancreatic transcription factors are linked to diabetes, including those functionally linked to the paired box transcription factor, Pax6. *Pax6* deletions in adult mice cause rapid onset of classic diabetes, but the full spectrum of pancreatic *Pax6* regulators is unknown. Using a regulatory element discovery approach we identified two novel *Pax6* pancreatic cis-regulatory elements in a poorly characterised regulatory desert. Both new elements, PE3 and PE4, are located 50 and 100 kb upstream, interact with different parts of the Pax6 promoter and nearby non-coding RNAs. They drive expression in the developing pancreas and brain, and code for multiple pancreas related transcription factor binding sites. PE3 binds CTCF and is marked by stem cell identity markers in embryonic stem cells, whilst a common variant located in the PE4 element affects binding of Pax4, a known pancreatic regulatory, altering *Pax6* gene expression. To determine the

ability of these elements to regulate gene expression synthetic transcriptional activators and repressors were targeted to PE3 and PE4, modulating *Pax6* gene expression, as well as influencing neighbouring genes and lncRNAs, implicating the *Pax6* locus in pancreas function and diabetes.

## Introduction

Distal cis-regulatory elements are a major component of the mechanism asserting temporal and spatial patterns of gene expression. Understanding the function of regulatory elements has taken on new significance as they are increasingly linked to human phenotypic variation and complex disease phenotypes. However as approximately 90% of disease-associated risk alleles fall within non-coding regions, a major challenge is pinpointing target genes and understanding underlying mechanisms of dysregulation (1).

Pax6 is an evolutionarily conserved pleiotropic transcription factor with roles in development of the central nervous system (CNS), the eye, the olfactory system and is also critical for pancreas development and hormone production from endocrine secretory cells (2, 3). In humans the congenital eye malformation aniridia is characterized by haploinsufficiency for the PAX6 protein and studying this condition has enabled the identification of large cis-regulatory regions controlling *Pax6* expression (4). In the pancreas multiple complex transcriptional networks utilise homeo- and paired-domain-containing transcription factors (such as Nkx2.2, Nkx6.1, Pdx1, Pax4, Isl1, and Pax6) (5) which are vital for coordinating the differentiation of progenitors to mature pancreatic cells (6) and directly regulate many pancreatic target genes (5, 7). Pax4 is another well-studied pancreatic transcription factor and is expressed early in pancreas development, where it is essential for specification and maintenance, as mice null for *Pax4* have a severe diabetic phenotype (8). Pax4 is also important for adult β-cell function and is linked to human pancreatic disease; mutations and common risk alleles in *Pax4* have been linked to Type1 diabetes (T1D) and Type 2 diabetes T2D (9–11).

Diabetes is caused by loss or dysfunction of the insulin secreting pancreatic β-cells, with autoimmune loss causing T1D, while in T2D Insulin secretion is defective, which in turn brings about an imbalance in glucose homeostasis and insulin resistance (9). Pax6 is required for embryonic stem cell differentiation to neural lineages consistent with its critical role in neural development. It is also expressed in the early pancreatic bud and is necessary for insulin homeostasis in the adult pancreas (12, 13); Pax6 deletion rapidly leads to classical diabetes and weight loss (13). A common regulatory variant (rs11603334G>A) for fasting pro-insulin levels (14, 15) is located in the *ARAP1* promoter, a regulator of *PAX6* whilst a genome-wide association to BMI ($p \leq 5.0\times10^{-7}$) is found upstream of the *PAX6* gene (16). Consistently a study of aniridia patients with heterozygous *PAX6* mutation found glucose intolerance characterized by impaired insulin secretion in all patients, demonstrating the endocrine pancreas is sensitive to levels of PAX6 (17). Similarly genome wide profiling of *cis*-regulatory networks in islet cells have shown an enrichment of T2D SNPs in islet specific enhancers, which themselves bind islet transcription factors (18).

Within the 11p13 locus the most critical gene encodes PAX6, which has a large regulatory domain with multiple long range elements, many of which reside within introns of its neighbouring gene, *ELP4*, including the downstream regulatory region (DRR), a complex enhancer cluster of tissue specific hypersensitive sites (19) (Figure 1A). Of importance for pancreatic function the pancreas and ectoderm enhancer cluster (P/EE) drives α and β-cell specific expression during development and after birth (20) and a pancreas-specific regulatory element (PE2) that drives stable endocrine pancreas expression during development and into adulthood (21). Targeted deletions of P/EE and PE2 regions reduce *Pax6* expression in the pancreas (22), but do not abolish it, we therefore hypothesised that other, as yet unidentified, pancreas-specific *Pax6* regulatory elements exist, and would be functionally conserved between human and mouse and reveal novel insight into *Pax6* pancreatic function and regulation.

To identify and functionally characterize novel regulatory elements in the *PAX6* regulatory domain we analysed human pancreatic tissues, mouse pancreatic β-cells (β-TC3) and ES cells as tractable experimental systems. Two elements were identified that associated with chromatin marks indicative of regulatory function in human and mouse pancreatic cells, called PE3 and PE4, they acted as regulatory elements in mouse reporter transgenics, revealing a neural and pancreatic expression pattern. Functional characterization of PE3 showed it was bound by CTCF and engaged the *Pax6* gene via regulatory looping over a 50 kb region, in both pancreatic cells and in mouse embryonic stem cells (mESC), whilst PE4 showed a more complex interaction profile in pancreatic cells interacting with PE3, the *Paupar* lncRNA and the *Pax6* gene. Within the PE4 element, a common variant, rs7943160G>C, is positioned in a vertebrate conserved PAX4 binding motif, and alters reporter expression in a PAX4 dependent manner, linking two pancreatic transcription factors. Finally, we demonstrated the importance of these cis regulatory sites by recruiting transcription activator-like (TAL) effectors fused to transactivator or repressor domains to the PE3 and PE4 elements to modulate expression of PAX6 and surrounding genes.

## Results

### Identification of novel human and mouse conserved pancreatic regulatory elements

In contrast to the downstream region of *PAX6,* which has a significant role in disease aetiology of aniridia patients, the upstream region towards *RCN1* is less well studied; it only includes the E-200 element and E-55 cluster, implicating this region as a regulatory "desert" even though it is known to contain a number of evolutionarily conserved sites (23, 24). As cis regulatory elements are key for modulation of gene expression through transcription factor binding, and are increasingly linked to complex disease phenotypes, we set out

to identify and characterise novel *PAX6* regulatory elements specific for pancreatic expression, due to its roles in pancreatic development and T2D. Reasoning that novel regulatory elements would be marked by enhancer specific histone modifications, we mined histone H3K27ac and H3K4me1 ChIP data sets from human primary pancreatic islet tissue (Human Epigenome atlas). Analysis of the region upstream of the *PAX6* gene suggested it might harbour a number of putative *PAX6* regulatory elements. Peak calling was used to identify pronounced ChIP signal enrichment, marking 5 discrete peaks in the upstream region (Figure 1B), labelled A-E, as putative novel cis regulatory elements. Analysis of transcriptome data from primary purified human pancreatic beta cells (25), confirmed high *PAX6* expression but unexpectedly two long noncoding RNAs (lncRNAs), *PAUPAR* and *PAX6-AS1*, located upstream of Pax6 were also expressed in β-cells (Fig 1B). Quantification across 6 individual primary β-cells (Table S1) indicated that *PAUPAR* was consistently expressed (mean FPKM 5.7) and aligned with the islet H3K27ac signal. The second lncRNA *PAX6*-AS1 is analogous to mouse *Pax6OS1* (26) (mean FPKM 10.2) and together this expression data suggests that *PAX6* locus lncRNAs may have a role in pancreatic cells.

Although this data hinted at the presence of regulatory elements, H3K27ac enrichment alone is not definitive for their identification. Important regulatory elements are likely to be well conserved in sequence and function between mammalian species, so putative regulatory marks in mouse cells were investigated. The strict tissue specificity of *Pax6* expression requires that ChIP enhancer profiling must be performed in a suitable cell type to identify appropriate active regulatory elements. Few pancreatic β-cell lines are available but mouse pancreatic β-TC3 are used in many studies and express *Pax6* at a high level (27), so provide a suitable model. ChIP-on-chip for the active enhancer modification H3K27ac was performed and mapped across the *Pax6* region using custom tiling arrays in β-TC3 cells (Fig 1C). The H3K27ac modification marked the *Pax6* promoter region and promoters of the adjacent ubiquitously expressed genes, *Elp4, Immpl1* and *Rcn1*. H3K27ac enrichment extended beyond the *Pax6* promoters, upstream where known *Pax6* pancreatic elements P/EE and PE2 are located and over the *Pax6* intron 7 enhancer cluster, 7CE1-4. Strikingly the mouse β-TC3 H3K27ac signal showed multiple novel H3K27ac enriched regions in the *Pax6* upstream region (labelled elements E-52, E-95, and E-120; 52kb, 95kb and 120kb 5' from the mouse P0 promoter respectively). There was high concordance between the β-TC3 and the human pancreatic islet data; human H3K27ac peaks A and B share core sequences with the mouse H3K27ac peaks E-52 and E-120. Previously, only the P/EE and PE2 elements were characterized as PAX6 pancreatic enhancers (20–22); thus identification of these new elements significantly expands the repertoire of regulatory regions with a

potential role in controlling *PAX6* expression in the pancreas. High resolution analysis of mouse and human DNaseI ENCODE data confirmed that these elements were discrete (data not shown), so were named as putative cis-regulatory Pancreatic Element 3 (PE3; corresponding to human peak A and mouse E-52), and PE4 (corresponding to human peak B and mouse E-120) (Fig 1).

We hypothesized that gene expression patterns as well as cis regulatory elements would be conserved between mouse and human pancreatic cells, so RNA-seq in β-TC3 cells was performed (Fig 1B). This revealed high *Pax6* (mean FPKM 110) and *Paupar* and *Pax6Os1* lncRNA (mean FPKM 2.2 and 6.2) expression (Table1), as seen in human pancreatic samples. To further assess the cell type specificity of these putative elements, we analysed Epigenome Atlas H3K4me1 ChIP-seq data across 48 distinct human tissues (Fig S1). There was visible enrichment of H3K4me1 signal over the PE3 region in 45% of tissues assayed, while PE4 only showed enrichment in pancreatic islet tissue samples (PI_13 and PI_27), suggesting it was pancreas-specific.

## PE3 and PE4 elements transactivate expression in mouse reporter transgenics

To assess the spatio-temporal characteristics of the novel PE3 and PE4 regulatory elements we generated LacZ reporter transgenic mice (20, 21). The putative regulatory elements were cloned into a *Hsp*68 minimal promoter–LacZ reporter construct (4) and used to generate transient transgenic embryos and stable lines, that were analysed at different developmental stages (Fig S2, Figure 2). For the PE3 element, 6 independent E11.5 dpc transient transgenic embryos were analysed before obtaining two stable lines, one of which was studied in detail across multiple developmental stages. Consistent staining between transient embryos and stable lines was observed in regions of the pancreas, brain and neural tube (Figure 2A-D). Midbrain expression was consistently observed for the PE3 element which is not a site of *Pax6* expression, however such occurrences of ectopic expression sites are not unusual in reporter transgenics when the element is not in its correct genomic context (28).

At E10.5 the PE3 element showed staining in the pancreas primordium, parts of the CNS including along the neural tube in the basal plate and dorsal root ganglia, as well as in the ventral midbrain and hindbrain (Fig 2A, B). At E11.5 staining was visible in the midbrain and hindbrain, whilst opening up the body cavity revealed further pancreatic staining (Fig 2C,D). During later stages of development expression became restricted to specific regions of the CNS, in particular the lateral olfactory tracts, midbrain, cerebellum and regions of the hindbrain (Figure S2A, B).

For the PE4 element three stable lines were obtained and consistent expression observed in a more restricted manner in the pancreas and specific regions of the developing brain. The PE4 element showed a more diffuse staining pattern in the CNS at E11.5 (Fig 2E), but some staining could be seen in the centre of the pancreas primordium (Fig 2F). At E17.5 the pancreas showed strong staining in islet like cells (Fig 2G), consistent with the reported *Pax6* expression pattern (29). Also at E17.5 the lateral olfactory tract, cerebellum and focal regions of the hindbrain showed expression (Fig 2H, Fig 2C).

These results support PE3 and PE4 to be novel tissue-specific regulators driving reporter expression in embryonic pancreas and neuronal tissues. PE3 has a broad *Pax6* expression pattern across multiple stages of development, while PE4 is more restricted to the developing pancreas.

## Novel *Pax6* regulatory elements show sequence conservation, transcription factor binding and encode putative regulatory SNPs

To directly compare between the human and mouse locus, sequence conservation at the putative PE3 and PE4 elements along with conserved transcription factor binding motifs within their sequences was examined to identify pathways and potential regulators. The PE3 element contained a 474 bp block with 58% sequence identity between human and mouse (Fig S3A), with a fully conserved 38 bp core across multiple mammalian species (data not shown). Transcription factor binding motifs were identified within the element for the Sox and Oct genes (30). Gene expression analysis of these transcription factors using the Human Protein atlas (31), revealed that 9 were expressed (FPKM >1) in adult pancreas, suggesting they are candidate transcription factors for binding to this element (Table S2). Based on identification of pluripotency motifs we next investigated published ChIP-seq data over the *Pax6* region in mESC and noted that the PE3 element and neighbouring E-55 elements were associated with active H3K27ac and H3K4me1 enhancer modifications, the p300 transcriptional co-activator as well as pluripotent transcription factors Sox2, Oct4 and Nanog (Fig 3A) (32). Sox2 and Oct4 have been shown to bind the *Pax6* promoter in mESC and displace nucleosomes over the region (33). Consistently, *Pax6* is one of a group of transcription factors which are bivalently marked in ESC by H3K4me3/H3k27me3 ready for activation upon differentiation (34). In addition to these pluripotency factors sequence analysis using the CTCF binding Site Database prediction tool (35) identified a conserved CTCF binding site in PE3 (Fig S3A) and ChIP in mouse β-TC3 cells confirmed this site binds CTCF in vivo (Fig S3B). Together this data revealed PE3 is a novel distal site of active transcriptional regulation of *Pax6* in mESC, which binds important pluripotency transcription factors and is occupied by a key regulator of 3D chromatin organisation (36).

Alignment of human and mouse sequences for the PE4 element identified a 802 bp fragment with 74% sequence identity between human and mouse (Fig S3C), containing multiple conserved transcription factor binding motifs (Table S3), including motifs for the important pancreas developmental regulators Nkx2.2 and Pax4 (8, 37). Furthermore, gene expression analysis showed that 19 out of 28 transcription factors with putative recognition motifs in the element are expressed in adult pancreas (Table S3) and indicate that PE4 is a more pancreas specific element than PE3.

Cis-regulatory elements that are expected to have a major role in common disease phenotypes are likely to harbour common variants within the population that could modulate their regulatory activity. To address whether any of the common SNPs found in PE3 and PE4 could disrupt transcription factor binding we scanned the human elements for SNPs embedded within transcription factor motifs. PE3 harbours two SNPs rs11031498 and rs11031499 but these did not coincide with any transcription factor binding sites. In contrast, rs7943160G>C within PE4, is a common variant in the population (Fig S4) which overlapped three transcription factor binding motifs: KAISO, MZF1 and Pax4 (Table S3).

## A common variant in PE4 alters regulatory element reporter activity

Pax4, an important regulator of pancreatic development, is expressed in adult α and β-islet cells (10). As PAX4 has been linked to both T1D, T2D (9, 11) and islet function we investigated variants in its binding motif in PE4. SNP rs7943160 altered position 29 of the 30 bp PAX4 motif (MA0068.1) (Figure 3B), with the motif having a stronger match for the ancestral C base over the common G variant. Multi-species sequence conservation analysis of PE4 revealed the PAX4 motif was seen across 59 mammals, with strong conservation of a C base in the aligned position of interest (94.8% of species) (Figure S5). This suggested it is an important nucleotide position and thus a candidate for a common genetic variant that would alter PE4 regulatory function and *PAX6* expression. To evaluate PAX4 binding to PE4 β-TC3 cells were transfected with a FLAG tagged version of human PAX4 (Fig 3C); FLAG-PAX4 ChIP signal over mouse PE4 was highly enriched compared to control regions (Figure 3D), so demonstrated that PAX4 can bind to the PE4 element.

As the PAX4 motif in PE4 contains a variant that may alter PAX4 binding affinity, the effect of the rs7943160G>C variant on the activity of the human PE4 element was assessed using a dual luciferase reporter assay in β-TC3 cells. The conserved region of the human PE4 element, with either the rs7943160 G or C variant, was cloned upstream of a minimal promoter driving luciferase (Fig 3E) and transfected into β-TC3 cells. Both variants of the PE4 element showed a decrease in luciferase signal compared to the empty vector,

demonstrating repressive behaviour. Importantly there was a striking and significant change in luciferase signal between the PE4(G) and PE4(C) variants ($p < 0.001$). To assess the effect of PAX4 on the variant binding sites, human *PAX4* cDNA was co-expressed with luciferase constructs in β-TC3 cells (Fig 3E). As PAX4 is a transcriptional repressor of pancreatic genes (38), based on the predicted effect of the variants on the strength of the PAX4 motif *in silico* (Fig 3A), we hypothesized that the G variant element would have lower binding affinity for PAX4 resulting in decreased repression and an increased luciferase signal. Both PE4 variants showed more repression when PAX4 cDNA was co-expressed and there was a significant difference between the two alleles ($p < 0.0001$), which was exaggerated compared to non-PAX4 cDNA transfected samples (Fig 3E). Together this indicated that a common regulatory variant in PE4 can alter regulatory element function and *PAX6* gene expression.

## Chromosome conformation capture reveals PE3 and PE4 regulatory looping to *Pax6*

Looping interactions have been proposed as a mechanism for regulatory elements located 10-100 kb's from target sites to interact and influence gene activity. Previously chromosome conformation capture (3C)-qPCR has been used to analyse interactions at multiple loci, so we selected this approach for characterising regulatory interactions around *Pax6* from the PE3 and PE4 elements. We first investigated PE3 as it binds CTCF (Fig S3C) and CTCF's role in coordinating gene regulatory looping is well established (36). PE3 was used as an anchor site from where relative interaction frequency was assayed at regular intervals across a panel of primers covering the 74 kb genomic landscape from PE3 to the *Pax6* gene. The PE3 element showed increased cross-linking frequency over the TSS of the *Paupar* lncRNA, whilst the signal increased further at 47 kb away from the anchor, over a region of high H3K27ac upstream of the promoter (Figure 1B), at the location of two known pancreatic regulatory elements P/EE and PE2 and then peaked over the *Pax6* P0 and P1 promoter fragments (Fig 4A). This data showed that PE3 is a regulatory element which sits in spatial proximity to active *Pax6* promoters in β-TC3 cells. The binding of pluripotency transcription factors at both the PE3 enhancer and promoter may facilitate *Pax6* being maintained in a poised state (39). As such, we hypothesised that the PE3 element would be important for initial *Pax6* gene activation, and maybe involved in priming the gene for subsequent expression in specific embryonic lineages such as neuroectoderm. Consistently 3C-qPCR in mouse embryonic stem cells revealed a more discrete interaction profile with signal peaking over background 3 kb upstream of the P0/P1 *Pax6* promoter, but reduced compared to that seen in β-TC3 cells (Fig 4A).

We next hypothesized that PE4 would interact with the *Pax6* promoter and gene body and these interactions would be detectably higher than in mESC cells where the element is not marked as active by H3K27ac. As predicted β-TC3 cells showed a complex profile with substantially higher relative crosslinking at both PE3 and multiple regions of the *Pax6* gene promoters and gene body than in mESC (Fig 4B), though the mESC sample did show increased signal over the PE3 and *Pax6* gene, suggesting the whole locus maybe in a conformation permissive for further activation. In both β-TC3 and mESC PE4 relative crosslinking frequency was higher over a broad region encompassing the E-55 and PE3 elements (Fig4B), consistent with this region being H3K27ac positive in mESC and β-TC3 cells (Fig1B, 2A). Interesting the two elements showed very distinct interaction patterns to the Pax6 promoter indicating that regulatory elements have a high degree of specificity for targeting associations between specific regulatory regions.

## Recruitment of synthetic transcription factors to PE3 and PE4 regulates gene transcription

Characterization of cis regulatory elements outside of the native genomic environment can only provide an incomplete picture of their function. TAL (transcription activator-like) effectors are a class of proteins that have been used to modulate transcription, epigenetic states (40, 41) and nuclear organisation (42, 43). As regulatory elements are landing pads for transcription factor binding, which transfer signals and factors onto gene promoter sequences, we reasoned that targeting a regulatory elements with a transactivator could be a useful approach to understand its specificity and behaviour. To examine the ability of the PE3 and PE4 elements to influence native *Pax6* expression we used a synthetic transcription factor modulation system (Figure 5A) using TAL effectors targeted to cis regulatory elements coupled to either a VP64 transcriptional co-activator (42) or a SID4X transcriptional repressor (40, 44) (Figure S3A,C). We reasoned that transcriptional modulation by synthetic transcription factor recruitment at the distal regulatory sites would affect *Pax6* transcription if the elements were *bona fide* cis regulatory elements for the gene.

As the PE3 element is a site for Sox2 and Oct4 binding (Fig 3A) TAL effector constructs were tested in mESC. Constructs were transfected into cells, enriched and RNA purified for qRT-PCR. TAL-VP64 targetted to PE3 caused a significant 1.6 to 2 fold increase in *Pax6* expression (Fig 5B) whilst expression of the TAL-SID4X (Fig 5A) repressor promoted a similar reduction in *Pax6* expression. The neighbouring genes *Elp4* and *Immpl1* (~235 kb downstream from *Pax6* promoter) showed an increase in expression whilst the upstream gene *Rcn1* (~270 kb away) showed no significant change in expression, nor did the control *Oct4* gene (Fig 5B). In contrast SID4X recruitment reduced expression of both the *Pax6* and *Rcn1 genes.*

To test the activity of the PE3 and PE4 elements in a pancreatic environment, constructs were expressed in β-TC3 cells. qRT-PCR for *Pax6* revealed significant transcriptional upregulation (2.5 to 3.4 fold) (Figure 6C) after VP64 recruitment, demonstrating PE4 can act over 120 kb to affect *Pax6* expression. As the H3K27ac signal in β-TC3 cells (Figure 1B) extends ~8 kb upstream of the P0 promoter over the lnRNA *Paupar* (45), we hypothesised that *Paupar* might be active in β-TC3 and modulated by factor recruitment to PE4. *Paupar* was indeed expressed to a similar level as *Pax6* and showed a strong and significant four-fold increase in expression upon VP64 recruitment (Fig 6B). As for mESC neighbouring genes were also influenced by VP64 recruitment, particularly downstream of the gene, consistent with promoters interacting locally with each other (46), but suggesting that *Rcn1* might be too far away to be strongly influenced by *Pax6* elements. SID4X recruitment did not have the repressive effect on gene expression seen in mESC, but surprisingly induced a small increase in expression at *Pax6*, *Paupar* and *Immpl1*. VP64 recruitment to PE3 had an even greater effect on *Paupar* expression than PE4 activation (Fig 5D) so we speculated that PE3 and PE4 function might be conserved in human pancreatic beta cells and regulate *Paupar* lncRNA in a similar manner.

To further investigate the co-expression of Pax6 with surrounding genes the correlation between gene expression across 23 RNA-seq samples in mouse tissues and cell lines (Fig S6A) was analysed. Pax6 expression was well correlated to downstream neighbouring genes, *Pax6Os1* (0.73) and *Elp4* (0.53), and showed little or a negative correlation with genes further upstream, *Rcn1* (0.004) and *Wt1* (-0.73). Consistently Pax6 is in a shared topologically associated domain (TAD) with surrounding genes as seen by HiC data visualisation in mESC (Fig S6B).

## Discussion

The *Pax6* locus has a densely packed regulatory landscape with more than 15 distinct regions of regulatory activity and many more predicted sites. Using complementary approaches, we identified two novel *Pax6* regulatory elements that are conserved in sequence and function between human and mouse. These novel elements drive diverse tissue specific developmental expression patterns, at multiple stages of mouse development, and act as key signal input sites for gene expression modulation in pancreatic and embryonic stem cells.

Transcription factor binding is a critical property of regulatory elements; multiple putative binding sites including CTCF and Pax4 were identified and characterised at PE3 and PE4, respectively (Fig S3), along with components of the ESC transcription factor network, Sox2, Oct4 and Nanog, (Fig 2A). CTCF is often

297 considered as a chromatin looping factor; consistent with this we showed PE3 interacted with the *Pax6*
298 promoter region in mESC. We propose that PE3 also has a role in recruiting the transcriptional apparatus
299 to *Pax6* in differentiated β-TC3 pancreatic cells, and maybe an element important for general Pax6 gene
300 activation in multiple tissues including the pancreas, consistent with its broad developmental expression
301 pattern (Fig 2). Recruitment of preinitiation complex components to distal elements has been described at
302 early stages of transcription at globin genes (47, 48), and likely facilitates efficient transfer of factors ready
303 for later differentiation. Using synthetic transcription factors we confirmed that PE3 can transfer both
304 activator and repressive signals over more than 45 kb to modulate *Pax6* gene expression. Interestingly we
305 were able to repress *Pax6* gene expression in mESC via TAL recruitment, but not β-TC3 cells using SID4X,
306 this could be due to the mechanism of action of the Sin3 domain, which likely acts via Hdac1/2 and histone
307 deacetylation (49). The correct combinations of factors may not be bound or available in β-TC3 cells or the
308 factors involved in transcriptional activation may override these signals. ESC transcription may also be more
309 plastic and more easily modulated; early establishment of enhancer activity and a poised bivalent gene
310 expression state in ESC is proposed as a means of efficient activation of target genes on differentiation, and
311 bivalent marks are focused on homeodomain transcription factors such as Pax6 (39). This is supported by
312 Sox2 binding to neural lineage-specific genes in advance of gene expression, which are then activated during
313 differentiation (50). We propose that early activation via regulatory looping of PE3 (Fig 4A), may transfer
314 ESC transcription signals which can specify ESC identity (39) and likely poises *Pax6* for rapid activation as a
315 transcription factor in neural differentiation.

316 The PE4 element has a more tissue restricted expression pattern than PE3 (Fig 2), and in a pancreatic cell
317 line model PE4 was regulated by the well-characterised pancreatic regulator, PAX4, through direct binding
318 to the element (Fig 3). Furthermore, the human PE4 element encodes a single nucleotide common variant
319 in a conserved PAX4 motif (Fig 3B), which modulates its reporter activity in a PAX4 dependent manner.
320 This demonstrates that a *PAX4*/*Pax6* regulatory network can be modulated by sequence variants found in
321 the population (Fig S4). Both PAX4 and PAX6 have a well-established link to diabetes phenotypes (9, 12, 13,
322 17). PAX4 is linked to the susceptibility of β-cells to apoptosis, leading to diabetes (9) and is found to be
323 significantly differentially expressed within a T2D cohort of adult islets (10). Similarly, mutations in a
324 number of transcription factors for islet function including PAX4, PAX6 (3), HNF1 A(51), Nkx2.2 (52) , and
325 SOX4 (53), cause diabetes or diabetes-like phenotypes in human and mouse. Unsurprisingly genome-wide
326 association studies for T2D are now revealing common variants at islet transcription factor loci (54). Pax6

regulates multiple β-cell specific genes (insulin 1 & 2, *Pdx1*, *GLUT2*, Nkx6.1) (7), so subtle dysregulation of a master regulator may cascade through transcriptional networks to cause downstream phenotypic effects. Such a scenario fits with a study suggesting a link between sequence variation in pancreatic islet cell enhancers and T2D (18). Therefore, the identified functional variant in the PE4 element might play a role in modulating gene function and contribute to tissue specific human phenotypic variation. As Pax6 expression in the adult pancreas has been shown to be important for islet maintenance and function we propose a model where subtle variations in the tight regulation of Pax6, via a cis-regulatory mechanism, interact with other genetic and environmental risk factors to affect T2D disease risk.

Synthetic transcription factor recruitment to a cis regulatory elements is an important tool to test and functionally dissect element activity and specificity. Using synthetic transcription factors (Fig 5) we showed that PE3 and PE4 could regulate *Pax6* expression in pancreatic β-TC3 cells. We also found TAL recruitment influenced neighbouring gene expression to suggest a complex interplay between locally interacting or regulating genes. This could be directly through shared element interactions with neighbouring genes in the same TAD or via gene clustering in a hub influencing one another's expression (55). This questions the idea of regulatory elements controlling single target genes and is consistent with experiments using the sleeping beauty transposon as a regulatory sensor which shows much of the genomic region around target genes are permissive to regulatory signals (56).

The *Pauper* lncRNA was upregulated by VP64 recruitment to PE3 and PE4 elements, this combined with interaction data that showed high relative crosslinking frequency over the *Pauper* gene indicates it is a target of these elements. *Pauper* was previously identified as a CNS specific lncRNA expressed from the *Pax6* locus, that was able to bind multiple regulatory elements in cis and trans, and linked to intraocular tumours (45, 57). Using shRNA mediated downregulation of *Paupar,* Vance et al. (45) showed it could transcriptionally regulate *Pax6,* whilst *Pax6* knockdown did not affect *Paupar* levels, suggesting the mechanism was not via a Pax6 auto-regulatory affect. Our data indicates that *Pax6* and *Paupar* are coupled at the level of transcription, potentially via shared elements or linked promoter activity. In a recent study of human pancreatic β-cell lnRNAs, Akerman et al. (58) found cell type specific lnRNAs play an important role in transcriptional regulation of multiple important pancreatic transcription factors acting both in cis and in trans, and were significantly altered in type 2 diabetic donor islets. Of particular interest the downregulation of the lnRNA neighboring the *PDX1* transcription factor gene, *PLUTO*, altered 3D enhancer interactions with *PDX1*, and could suggest a shared mechanism of cis-regulator function with *Paupar/Pax6*. Our data

reveals complex patterns of transcription factors and binding motifs at novel pancreatic cis-regulatory elements. These tune tissue-specific *PAX6* gene expression, can be modulated by common genetic variants, and further implicate the *Pax6* locus in pancreas function and diabetes.

## Materials and Methods

### Cell lines

β-TC3 cells were isolated from a mouse insulinoma (27) and were cultured in Dulbecco's Modified Eagle Medium (ThermoFisher) supplemented with 10% fetal calf serum and 1% Penicillin-Streptomycin at 37°C in 5% CO2. Mouse OS25 embryonic stem cells where cultured in Glasgow's MEM (ThermoFisher) supplemented with 10% fetal calf serum, 1% Penicillin-Streptomycin, 1% MEM Non-essential Amino Acid Solution (Sigma), 1mM Sodium Pyruvate (Sigma), recombinant LIF and 0.01 mM 2-Mercaptoethanol (Gibco), at 37°C in 5% CO2 using standard techniques.

### Chromatin Immuno-precipitation

H3K27ac (Abcam, ab4729) or CTCF ChIP (Cell signaling, D31H2 XP Rabbit mAb #3418) was performed as described previously (19) using Protein G Dyna beads (ThermoFisher Cat#10003D) with two biological replicates. Primers where designed using Primer3 to PE3 regions with flanking control primers (List of primers). QPCR was performed using LightCycler® 480 SYBR Green I Master Mix (Roche Cat#04707516001) according to the manufacture's guidelines and using a LightCycler® 480 II, with primers at a final concentration of 0.5 μM. Ct values were used to calculate ChIP enrichment at each primer region vs 10% of Input DNA.

H3K27ac ChIPs were validated by qPCR before hybridisation to genomic microarrays (Nimblegen 720K) covering a 66 Mb region around *Pax6* (Chr2:75,000,000-141,000,000). ChIP and Input samples where amplified (GenomePlex, Sigma), purified (QIAquick, Qiagen) and labelled (NimbleGen Dual-Colour Labelling Kit, Roche Cat. 06370250001). ChIP samples (Cy5) and Input samples (Cy3) were hybridized using a NimbleGen Hybridization and Sample Tracking Control Kit (Roche Cat. 05993776001) according to the manufacturer's instructions. Slides were washed (NimbleGen Wash Buffer Kit, Roche Cat. 0558450700) and scanned at 2 μm resolution on a MS 200 Microarray Scanner (Nimblegen). Images were processed using NimbleScan (version 2.5); Ringo (Bioconductor) was used for pre-processing, normalisation, combining replicates and peak calling of ChIP-chip data. Data was further processed in R by applying a running median (500 bp) and visualised on the UCSC genome browser.

Human PAX4 cDNA (Transgenomics) was PCR amplified (Supplementary Table 5) and cloned into pcDNA5/FRT/TO/3xFlag with *Hind*III/*Xho*I. pcDNA5/FRT/TO3xFLAG was generated by ligating 3xFLAG (amplified from p3xFLAG-CMV-10(SIGMA); List of primers) into pcDNA5/FRT/TO at *Afl*II/*Bam*HI. PAX4-FLAG ChIP was performed as described as above with the following modifications. β-TC3 cells were transfected using Lipofectamine 2000 according to the manufacturer's protocol (Thermo Fisher) in Opti-MEM, using 12 μg of PAX4-FLAG construct per 10 cm dish and cells were harvested after 48 h. IPs were performed using mouse monoclonal anti-FLAG M2 (Sigma), Mouse IgG as control, and Sheep anti-Mouse IgG M-280 Dynabeads (Thermo Fisher). QPCR was performed on ChIP material with SYBR Select Master Mix (Thermo Fisher Cat# 4472908) on a LightCycler 480 II, standard protocol. To quantify ChIP enrichment primers were designed to the mouse PE4 element and control regions (List of primers), and enrichment calculated % Input.

## Luciferase reporter assays

β-TC3 cells were grown overnight in a 24 well culture plate. Cells were transfected with luciferase reporter constructs (as described below), human PAX4 or PAX6 cDNA (Transgenomics), and with Renilla luciferase pRL-TK (Promega) as an internal control. Luciferase assays were performed 48 h after transfection using a Dual-Luciferase Reporter Assay System (Promega). Relative luciferase activity was calculated by dividing Firefly Luciferase signal by Renilla Luciferase signal and normalising the resulting value to the relative luciferase activity of the negative control vector. Five biological replicates of each assay were performed. To analyse expression of PAX4, PAX6 and FLAG tagged constructs cell extracts were prepared in 4 x LDS sample buffer and analysed by protein gel electrophoresis (NuPage, Invitrogen) and Western blotting with anti-PAX4 (Abcam,135598), anti-PAX6 (AD1.5.6 and AD2.35) (59), anti-hnRNPU (Millipore, 05-1516) and anti-FLAG M2 (Sigma) primary antibodies. Luciferase reporter constructs were made by cloning the putative PE4 element (Chr11:31947517-31948318) into a minimal promoter pGL4.23 vector (Promega). Single base pair changes corresponding to the SNP variants were introduced into the putative PAX4 motif within the PE4 element by site directed mutagenesis using a mega-primer approach and confirmed by sequencing.

## Mouse Reporter transgenics

Mouse LacZ reporter transgenics were derived as described previously (4). Three founders were obtained for the PE3 element (chr2:105456479 -105456966), two of which showed an expression pattern consistent with previously obtained transiently expressing embryos. Line PE3Z-004 was selected as representative and multiple embryos from PE3Z-004 male x wild type female matings were analysed at three developmental stages; E10.5, E11.5, and E17.5. Two stable LacZ lines were established for the PE4 element (chr2:105390725-105393222; PE4-Z-011 and PE4-Z-029) and representative embryos were analysed at three developmental stages, as above. All animal experiments were approved by the University of Edinburgh ethical committee (approval ID TR-15-08) and performed under UK Home Office license number PPL 60/3785.

## RNA-seq

Two experimental replicates where generated with two T25 culture flask of β-TC3 cultured to ~70% confluency. Total RNA was prepared using the Qiagen RNeasy mini kit (Qiagen) with an on-column DNase I digestion using the Qiagen RNase-Free DNase set as the manufactures guidelines. Ribosomal RNA depletion was performed using the RiboMinus Eukaryote Kit for RNA-Seq (Life Technology) and RiboMinus Concentration module as manufactures' guidelines, with ribosomal depletion was confirmed by gel electrophoresis. Sequencing libraries where prepared using NEBNext mRNA Library Prep Master Mix Set (NEB), following the manufactures protocol, with all size selection performed using Agencourt AMPure XP beads (Beckman Cutler). Samples where multiplexed suing the NEBNext Multiplex Oligos for Illumina kit to barcode each sample and amplified for 11 cycles. The libraries were analysed using an Agilent bioanalyser high sensitivity DNA chip. The subsequent samples were processed by the Next Gen Sequencing facility at the *VU University Medical Centre,* Department of Pathology Amsterdam, using an Illumina Hi-Seq 2000 producing single end 50 reads. Reads where aligned to mm9 geneome index using TopHat v2[67] and processed with Samtools v1.6 . Ethical approval was received to re-analayais RNA-seq data from 6 primary human β-Cells from Nica et al 2013. For both human and mouse samples aligned bam files where processed with Subread v1.5 *feature counts* to generate FPKM scores (using total number of aligned of reads and gene length) against hg19 and mm9 RefSeq genes (60). Bedtools genome coverage tool (61) scaled to number of aligned reads was used to generate visualization of read distribution across the human and mouse Pax6 locus.

## Bioinformatic data and analysis

All genomic positions are hg19 or mm9. Primary human histone modification ChIP-seq data was downloaded from the Human Epigenome Atlas (Table S4). Peak calling was performed using MACS version 1.4.2, P value cut off 0.005 (62). Conserved transcription factor binding sites within aligned human and mouse sequences where identified with rVista (30). Human-mouse sequence alignment was performed using Clustal Omega. Candidate transcription factor motifs altered by SNPs were identified using HaploReg and motifs were validated using the Jasper database (63, 64). Spearmen correlation on neighbouring gene expression was performed in R using RNA-seq RPKM values from 23 mouse tissue and cell line data sets, 20 samples from (65), differentiated neurons from (50), and β-TC3 cell RNA-seq, were analysed with Geneprof tool (66), (see Table S4 for details of 23 mouse samples).

### 3C-qPCR

The 3C procedure was based on the methods adapted from (67). A none ligase control sample were run alongside experimental samples, and produced no enrichment in qPCR assays. Adherent β-TC3 and OS25 mESC were treated with trypsin/versene and resuspended as a single cell suspension in 10% FCS/PBS. Cells were counted and $1x10^7$ cells were used per 3C experiment. Cells were fixed in 10% FCS/PBS with 2% (v/v) formaldehyde (Sigma) for 10 min at room temperature. Glycine was added, cells centrifuged and lysed in cell lysis buffer [10 mM Tris-HCl, ph 7.5; 10 mM NaCl; 0.2% (v/v) Ipegal, PI]. Cells were centrifuged and resuspened in 500 µl 1.2x NEB 3. 3C samples were incubated with SDS at a final concentration of 0.2% for 1 h, before Triton X-100 was added to 2% for 1 h. 1000 U of concentrated *Bgl*II restriction enzyme (NEB Cat# R0144M) was added for overnight digestion at 37°C and 1200 rpm. Digestion efficiency was analysed on 5 µl pre and post digested template by gel electrophoresis. Restriction enzyme was deactivated by heating at 65°C for 25 min in 1.6% SDS. Ligation of 3C library was performed in a final volume of 7mls diluted in 1x T4 DNA ligase reaction buffer (NEB), with 1% Triton X-100. Ligation was performed at 16°C for 4 h at 300 rpm with 3.3 µl of high concentration DNA ligase (NEB Cat#M0202M). Cross-links were removed by heating at 65°C overnight with 300 µg of PK, followed by incubation at 37°C for 1 h with 300 µg RNase A. Samples were purified by two sequential phenol-chloroform extractions (Sigma), ethanol precipitation, resuspended in 150 µl of 10 mM Tris pH7.5.

The PE3 and PE4 anchor fragment primers where designed with both primer and probe to lie within 150 bp of the *Bgl*II cut site. The probes where dual labelled with a 5' 6-FAM fluorophore and a 3' BHQ1 quencher.

The variable fragment primer panel (primer list) were designed to lie within 100-150 bp of the target *Bgl*II restriction site using an *in silico* digest of the mouse genome across the *Pax6* regions (mm9 build). The panel of variable fragment primers was validated on a random ligation template (RLT, design below) using the anchor primers and constant probe primer to confirm each primer and probe efficiency across a range of DNA concentrations.

BAC DNA for the Pax6 locus (RP23-281P3, Chr2: 105,427,870-105,581,806) and PAC for the control *Ercc3* region (PAC 334G18) where used to generate a RLT for the region of interest, to produce values of the standard curves for each primer combination. BAC and PAC DNA was prepared used a standard alkaline lysis mini-prep method. The RLT was prepared (68), with equimolar amounts of BAC and PAC DNA digested with Bgl*II* as for the 3C protocol, phenol chloroform extracted, and precipitated. The DNA was ligated at high DNA concentrations to promote intra-molecular ligation events and purified. Standard curves were generated using the RLT on the primer panel as described by (68), with serial 5-fold dilutions of RLT used to produce standard curves for all variable primers with the probe and constant fragment primer. Digested and ligated genomic DNA was used to keep the sample DNA concentration in the qPCR reaction constant and the probe qPCR protocol was performed as 3C samples qPCR. Absolute quantification was performed using a LightCycler 480 II (Roche) using LightCycler 480 software to generate the standard curve efficiency, slope and intercept values. Four qPCR reactions were performed for each primer sample combination, using QuantiTect Probe PCR master mix standard protocol (Qiagen Cat#204341), the probe at a final concentration of 0.15 μM and primers at 0.5 μM, the LightCycler480 program described in Hagège et al. (67), detecting 6-FAM signal. Two independent biological experiments were performed each generating individual Ct values and calculations were performed as in (67) with each Ct value used to calculate a relative cross-linking frequency using the parameters of the standard curve, and then normalised to *Ercc3* background interactions to allow comparison between samples. The final relative cross-linking and SEM calculation was performed on the mean relative cross-linking across both experimental replicates and this was plotted as a function of the distance from the anchor probe to the test primer.

## TALEs

TAL (transcription activator-like) effectors (TALEs) were assembled using a modular assemble system ) (41) in which specific RVD DNA binding domain is constructed through assemble of three 4-mer modules and

one-3mer plasmid module, linearized with BsmBI and cloned into pTAL-VP64-eGFP(42). The PE3 and PE4 regulatory element target sequence where identified using the TAL effector nucleotide targeter 2.0 tool (69), and were targeted to a 16bp sequence in the core element sequence (PE4_TAL: TCCTCAGGCCATGCAT, chr2:105392464-105392479; PE3_TAL: TCGAGCTAATCCTCTT, chr2:105456674-105456689). To generate SID4X TALEs the VP64 sequence was removed using (BamH1/Nhe1 digestion) and replaced with the SID4X sequence (see primers).

Mouse ES and β-TC3 cells where seed at ~70-80% confluency in 60 mm cell culture dishes 24hr before transfection and transfected in Opti-MEM with Lipofectamine 3000 (ThermoFisher) using a standard protocol with 5 μg of plasmid, and incubated for 48hours (β-TC3) and 36hours (mESC) . Cells where washed in PBS, trypsinized and flow sorted for eGFP expression and GFP positive cells collected. Flow sorting was performed on a BD FACSAria2 SORP cell sorter by MRC HGU FACS facility, GFP data was collected using a 488nm excitation laser and a 525/50nm bandpass emission filter, and BDFACS Diva software version 6.1.3 was used for data collection. GFP positive cells were collected and RNA extracted using a RNeasy mini Kit (Qiagen), with GFP negative mock transfected cells as control, all RNA samples were digested with DNaseI on column (Qiagen). RNA concentration was measured and corrected for all samples and reverse transcribed to cDNA using SuperScript III (ThermoFisher) using a standard first strand synthesis protocol with Oligo(dT) primers (Promega), for two biological replicates. Real time qPCR was used to measure transcript levels using a LightCycler 480 II (Roche), with SYBR Select Master Mix (ThermoFisher) following the manufacturer protocol with primers at 0.5 μM. Gene specific primers were designed to Ensemble cDNA transcripts (List of primers), we were unable to design primers which could assay Pax6OS1. The log2 fold change calculated using the ΔΔCT method against mock transfection controls, with 18S as a housekeeping gene.

## List of Primers

*Pax4 ChIP qPCR*

| | |
|---|---|
| ChrX_Con | TTCTGGGGTTTGTGCATGTG(f) AGAGTAGAAGGACGGTATTGGT(r) |
| PE4_1 | TAGCCGGTGTTCCATTGTCT(f) GCTAGTGTTTAAACCGCTCCA(r) |
| PE4_2 | GCCAACCAGACAATCTTCAGT(f) GCTGGGCTGTAATTTGCTGA(r) |
| PE4_3 | GGAGCGGTTTAAACACTAGCA(f) GAGCTTCTCTGGCAGCCTT(r) |
| PE4_Con | ATGTGCAGCTATCCCCATGT(f) TGTGGAATGCTCAGCCCTAA(r) |
| Chr2_Con | GTGGCACATCACAAATGCTC(f) TCTCCAGTCTAACACTTGGCAAT(r) |

*CTCF ChIP qPCR*

| | | |
|---|---|---|
| Hum.PE4.1 | CGCTAGTTTCAATTTGGCTGT(f) TGAATAGCGGCAAAGATCCTG(r) |
| Hum.PE4.2 | TGAAGATACCTGGATGAAGCACT(f) TGTGAATGAATAGCGGCAAAGA(r) |
| Luc.IP.1 | TTCGACCGGGACAAAACCAT(f) ATCTGGTTGCCGAAGATGGG(r) |
| Luc.IP.2 | TTCGGCAACCAGATCATCCC(f) GTACATGAGCACGACCCGAA(r) |
| PE3_1 | CCAGTGCTCTGGGCTACAAT(f) AAGACGCCAGGAAGAGGATT(r) |
| PE3_2 | AATCCTCTTCCTGGCGTCTT(f) GGAAGGCTCTGTCCCTCTTT(r) |
| Con_+1kb | TGTTTTGGGGTCTCCTGAAG(f) CATGGACTAACAATGCTTCTCCT(r) |
| Con_-1kb | AAGAGGACTCAGCGAAACCA(f) TGACCTCATGCCAACTCATC(r) |
| *Taqman probes* | |
| PE3_probe | CCTCTTCTTTCACAGTGTTCCCTGA |
| PE4_probe | CCTGCCTTGAAAACTCTTCTCGTCTCTG |
| Ercc3_probe | CCAGACCAGAGAGCGGAGACC |
| *3C primers* | |
| 3C_PE4.F | GGTGAGTCTTTACATGTGGGG |
| 3C_1.F | TCCCACCATCCAAATTCTGT |
| 3C_2.F | GCCATCTCTCTAGCCCCTTT |
| 3C_3.F | CCTCTTCTGTCCAGGCTTTG |
| 3C_PE3.F | AGCAAATGTGTGACCGTGAG |
| 3C_4.F | ATCCACCCACCTCCTTATCC |
| 3C_5.F | CTCTTTCTGGTTTGCGGTAT |
| 3C_6.F | CTGTTCTTCCTCTGAAACCTG |
| 3C_7.F | ACAGTGGCACGTTGGATATG |
| 3C_8.F | TGTGTGCAAATGAAGGCTCT |
| 3C_9.F | TGACCTGCAAGAAGACACAGA |
| 3C_10.F | CGCTTTGATTCTAGCCAGAC |
| 3C_11.F | GTGTAATTGAGGGAAATGGAGTTGAA |
| 3C_12.F | GGCCAGTTTGACACACCTTT |
| 3C_13.F | CCCCAACCTTTGTACTCAGC |
| 3C_14.F | TCTTTTGCCCAGAGATGAGC |
| 3C_15.F | GGAAAGGCACTTGGAAATGA |
| 3C_16.F | CTGGTGACCATCCACTCTCC |
| 3C_17.F | TGAGAGGACCCATTATCCAGA |
| 3C_errc3_1.F | GGCTGAGAGTGATGCTGCTA |
| 3C_errc3_2.F | CGGTAAATCTCCTCCCAAAT |
| *Cloning* | |
| hPAX4 | TCGCAAGCTTATGAACCAGCTTGGGG(f) TGCCCTCGAGTTATTCCAAGCCATACAG(r) |
| 3xFLAG | ACGACTTAAGGGCGCGCCACCATGGACTACAAAGACC(f) GTCAGGATCCTCTAGAGTCGAC(r) |

| | | |
|---|---|---|
| SID4X | ACGAGGATCCGGCTCCGGGATGAACATCCAG(f) GTCAGCTAGCTCTACTGGGCAGCATAGAGG(r) |
| *TAL RT-qPCR* | |
| Pax6 ex6-7 | AGTTCTTCGCAACCTGGCTA(f) GTGTTCTCTCCCCCTCCTTC(r) |
| Pax6 ex4-5 | CGTGCGACATTTCCCGAATT(f) CTTGGCTTACTCCCTCCGAT(r) |
| Paupar | TGCTCTTCTGTCTAGGGTGC(f) AACTTCATCCAAAAGGCCGG(r) |
| Oct4 | CGAGAACAATGAGAACCTTC(f) CCTTCTCTAGCCCAAGCTGAT(r) |
| 18s | GTAACCCGTTGAACCCCATT(f) CCATCCAATCGGTAGTAGCG(r) |
| Immpl1.ex3_4 | GCTTTTCGACTTGCTGGCTA(f) TGTCGGCTAAGATTTTCTGCA(r) |
| Rcn1.ex2_3 | AGAATACAAGCAGGCCACCT(f) GCAGAAAGGCAGTGAACTCC(r) |
| Elp4.ex3_4 | CATGGCAGAAGGAATCATCA(f) GCTCTGGGGTTTTAGCACTG(r) |


## Acknowledgements

We would like to thank Lora Boteva, Jennifer Huffman, Chris Ponting, Veronica van Heyningen and Malcolm Dunlop for advice and critical reading of the manuscript.

## Funding:

This work was funded by the UK Medical Research Council and NG is an MRC Senior Fellow (MR/J00913X/1).


## References:


1. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J., *et al.* (2012) Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, **337**, 1190.

2. Ashery-Padan,R., Zhou,X., Marquardt,T., Herrera,P., Toube,L., Berry,A. and Gruss,P. (2004) Conditional inactivation of Pax6 in the pancreas causes early onset of diabetes. *Dev. Biol.*, **269**, 479–488.

3. St-Onge,L., Sosa-Pineda,B., Chowdhury,K., Mansouri,A. and Gruss,P. (1997) Pax6 is required for differentiation of glucagon-producing alpha-cells in mouse pancreas. *Nature*, **387**, 406–409.

4. Kleinjan,D.A., Seawright,A., Schedl,A., Quinlan,R.A., Danes,S. and van Heyningen,V. (2001) Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Hum. Mol. Genet.*, **10**, 2049–2059.

5. Collombat,P., Mansouri,A., Hecksher-Sørensen,J., Serup,P., Krull,J., Gradwohl,G. and Gruss,P. (2003) Opposing actions of Arx and Pax4 in endocrine pancreas development. *Genes Dev.*, **17**, 2591–2603.

6. Wilson,M.E., Scheel,D. and German,M.S. (2003) Gene expression cascades in pancreatic development. *Mech. Dev.*, **120**, 65–80.

7. Gosmain,Y., Katz,L.S., Masson,M.H., Cheyssac,C., Poisson,C. and Philippe,J. (2012) Pax6 Is Crucial for β-Cell Function, Insulin Biosynthesis, and Glucose-Induced Insulin Secretion. *Mol. Endocrinol.*, **26**, 696–709.

8. Sosa-Pineda,B., Chowdhury,K., Torres,M., Oliver,G. and Gruss,P. (1997) The Pax4 gene is essential for differentiation of insulin-producing beta cells in the mammalian pancreas. *Nature*, **386**, 399–402.

9. Brun,T. and Gauthier,B.R. (2008) A focus on the role of Pax4 in mature pancreatic islet beta-cell expansion and survival in health and disease. *J. Mol. Endocrinol.*, **40**, 37–45.

10. Bonnavion,R., Jaafar,R., Kerr-Conte,J., Assade,F., van Stralen,E., Leteurtre,E., Pouponnot,C., Gargani,S., Pattou,F., Bertolino,P., *et al.* (2013) Both PAX4 and MAFA are expressed in a substantial proportion of normal human pancreatic alpha cells and deregulated in patients with type 2 diabetes. *PloS One*, **8**, e72194.

11. Cho,Y.S., Chen,C.-H., Hu,C., Long,J., Hee Ong,R.T., Sim,X., Takeuchi,F., Wu,Y., Go,M.J., Yamauchi,T., *et al.* (2012) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.*, **44**, 67–72.

12. Hart,A.W., Mella,S., Mendrychowski,J., van Heyningen,V. and Kleinjan,D.A. (2013) The developmental regulator pax6 is essential for maintenance of islet cell function in the adult mouse pancreas. *PloS One*, **8**, e54173.

13. Simpson,T.I. and Price,D.J. (2002) Pax6; a pleiotropic player in development. *BioEssays News Rev. Mol. Cell. Dev. Biol.*, **24**, 1041–1051.

14. Strawbridge,R.J., Dupuis,J., Prokopenko,I., Barker,A., Ahlqvist,E., Rybin,D., Petrie,J.R., Travers,M.E., Bouatia-Naji,N., Dimas,A.S., *et al.* (2011) Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes*, **60**, 2624–2634.

15. Kulzer,J.R., Stitzel,M.L., Morken,M.A., Huyghe,J.R., Fuchsberger,C., Kuusisto,J., Laakso,M., Boehnke,M., Collins,F.S. and Mohlke,K.L. (2014) A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *Am. J. Hum. Genet.*, **94**, 186–197.

16. Wen,W., Zheng,W., Okada,Y., Takeuchi,F., Tabara,Y., Hwang,J.-Y., Dorajoo,R., Li,H., Tsai,F.-J., Yang,X., *et al.* (2014) Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index. *Hum. Mol. Genet.*, **23**, 5492–5504.

17. Yasuda,T., Kajimoto,Y., Fujitani,Y., Watada,H., Yamamoto,S., Watarai,T., Umayahara,Y., Matsuhisa,M., Gorogawa,S., Kuwayama,Y., *et al.* (2002) PAX6 mutation as a genetic factor common to aniridia and glucose intolerance. *Diabetes*, **51**, 224–230.

18. Pasquali,L., Gaulton,K.J., Rodríguez-Seguí,S.A., Mularoni,L., Miguel-Escalada,I., Akerman,İ., Tena,J.J., Morán,I., Gómez-Marín,C., van de Bunt,M., *et al.* (2014) Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.*, **46**, 136–143.

19. McBride,D.J., Buckle,A., van Heyningen,V. and Kleinjan,D.A. (2011) DNaseI hypersensitivity and ultraconservation reveal novel, interdependent long-range enhancers at the complex Pax6 cis-regulatory region. *PloS One*, **6**, e28616.

20. Kammandel,B., Chowdhury,K., Stoykova,A., Aparicio,S., Brenner,S. and Gruss,P. (1999) Distinct cis-essential modules direct the time-space pattern of the Pax6 gene activity. *Dev. Biol.*, **205**, 79–97.

21. Xu,P.X., Zhang,X., Heaney,S., Yoon,A., Michelson,A.M. and Maas,R.L. (1999) Regulation of Pax6 expression is conserved between mice and flies. *Development*, **126**, 383–395.

22. Carbe,C., Hertzler-Schaefer,K. and Zhang,X. (2012) The functional role of the Meis/Prep-binding elements in Pax6 locus during pancreas and eye development. *Dev. Biol.*, **363**, 320–329.

23. Ravi,V., Bhatia,S., Gautier,P., Loosli,F., Tay,B.-H., Tay,A., Murdoch,E., Coutinho,P., van Heyningen,V., Brenner,S., *et al.* (2013) Sequencing of Pax6 Loci from the Elephant Shark Reveals a Family of Pax6 Genes in Vertebrate Genomes, Forged by Ancient Duplications and Divergences. *PLoS Genet*, **9**, e1003177.

24. Bhatia,S., Bengani,H., Fish,M., Brown,A., Divizia,M.T., de Marco,R., Damante,G., Grainger,R., van Heyningen,V. and Kleinjan,D.A. (2013) Disruption of Autoregulatory Feedback by a Mutation in a Remote, Ultraconserved PAX6 Enhancer Causes Aniridia. *Am. J. Hum. Genet.*, **93**, 1126–1134.

25. Nica,A.C., Ongen,H., Irminger,J.-C., Bosco,D., Berney,T., Antonarakis,S.E., Halban,P.A. and Dermitzakis,E.T. (2013) Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. *Genome Res.*, **23**, 1554–1562.

26. Alfano,G., Vitiello,C., Caccioppoli,C., Caramico,T., Carola,A., Szego,M.J., McInnes,R.R., Auricchio,A. and Banfi,S. (2005) Natural antisense transcripts associated with genes involved in eye development. *Hum. Mol. Genet.*, **14**, 913–923.

27. Henseleit,K.D., Nelson,S.B., Kuhlbrodt,K., Hennings,J.C., Ericson,J. and Sander,M. (2005) NKX6 transcription factor activity is required for alpha- and beta-cell development in the pancreas. *Development*, **132**, 3139–3149.

28. Griffin,C., Kleinjan,D.A., Doe,B. and van Heyningen,V. (2002) New 3' elements control Pax6 expression in the developing pretectum, neural retina and olfactory region. *Mech. Dev.*, **112**, 89–100.

29. Sander,M., Neubüser,A., Kalamaras,J., Ee,H.C., Martin,G.R. and German,M.S. (1997) Genetic analysis reveals that PAX6 is required for normal transcription of pancreatic hormone genes and islet development. *Genes Dev.*, **11**, 1662–1673.

30. Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W217–W221.

31. Uhlén,M., Fagerberg,L., Hallström,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,Å., Kampf,C., Sjöstedt,E., Asplund,A., *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.

677 32. Whyte,W., Bilodeau,S., Orlando,D.A., Hoke,H.A., Frampton,G.M., Foster,C.T., Cowley,S.M. and
678 Young,R.A. (2012) Enhancer decommissioning by LSD1 during embryonic stem cell differentiation.
679 *Nature*, **482**, 221–225.

680 33. Sebeson,A., Xi,L., Zhang,Q., Sigmund,A., Wang,J.-P., Widom,J. and Wang,X. (2015) Differential
681 Nucleosome Occupancies across Oct4-Sox2 Binding Sites in Murine Embryonic Stem Cells. *PLoS*
682 *ONE*, **10**.

683 34. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M.,
684 Plath,K., *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic
685 stem cells. *Cell*, **125**, 315–326.

686 35. Ziebarth,J.D., Bhattacharya,A. and Cui,Y. (2012) CTCFBSDB 2.0: a database for CTCF-binding sites and
687 genome organization. *Nucleic Acids Res.*, **41**, D188–D194.

688 36. Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.

689 37. Sussel,L., Kalamaras,J., Hartigan-O'Connor,D.J., Meneses,J.J., Pedersen,R.A., Rubenstein,J.L. and
690 German,M.S. (1998) Mice lacking the homeodomain transcription factor Nkx2.2 have diabetes due
691 to arrested differentiation of pancreatic beta cells. *Development*, **125**, 2213–2221.

692 38. Petersen,H.V., Jørgensen,M.C., Andersen,F.G., Jensen,J., F-Nielsen,T., Jørgensen,R., Madsen,O.D. and
693 Serup,P. (2000) Pax4 Represses Pancreatic Glucagon Gene Expression. *Mol. Cell Biol. Res. Commun.*,
694 **3**, 249–254.

695 39. Boyer,L.A., Lee,T.I., Cole,M.F., Johnstone,S.E., Levine,S.S., Zucker,J.P., Guenther,M.G., Kumar,R.M.,
696 Murray,H.L., Jenner,R.G., *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic
697 stem cells. *Cell*, **122**, 947–956.

698 40. Konermann,S., Brigham,M.D., Trevino,A., Hsu,P.D., Heidenreich,M., Le Cong, Platt,R.J., Scott,D.A.,
699 Church,G.M. and Zhang,F. (2013) Optical control of mammalian endogenous transcription and
700 epigenetic states. *Nature*, **500**, 472–476.

701 41. Ding,Q., Lee,Y.-K., Schaefer,E.A.K., Peters,D.T., Veres,A., Kim,K., Kuperwasser,N., Motola,D.L.,
702 Meissner,T.B., Hendriks,W.T., *et al.* (2013) A TALEN genome-editing system for generating human
703 stem cell-based disease models. *Cell Stem Cell*, **12**, 238–251.

704 42. Therizols,P., Illingworth,R.S., Courilleau,C., Boyle,S., Wood,A.J. and Bickmore,W.A. (2014) Chromatin
705 decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science*, **346**,
706 1238–1242.

707 43. Benabdallah,N.S., Williamson,I., Illingworth,R.S., Boyle,S., Grimes,G.R., Therizols,P. and Bickmore,W.
708 (2017) PARP mediated chromatin unfolding is coupled to long-range enhancer activation. *bioRxiv*,
709 10.1101/155325.

710 44. Cong,L., Zhou,R., Kuo,Y.-C., Cunniff,M. and Zhang,F. (2012) Comprehensive interrogation of natural
711 TALE DNA-binding modules and transcriptional repressor domains. *Nat. Commun.*, **3**, 968.

45. Vance,K.W., Sansom,S.N., Lee,S., Chalei,V., Kong,L., Cooper,S.E., Oliver,P.L. and Ponting,C.P. (2014) The long non-coding RNA Paupar regulates the expression of both local and distal genes. *EMBO J.*, **33**, 296–311.

46. Mifsud,B., Tavares-Cadete,F., Young,A.N., Sugar,R., Schoenfelder,S., Ferreira,L., Wingett,S.W., Andrews,S., Grey,W., Ewels,P.A., *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.

47. Vieira,K.F., Levings,P.P., Hill,M.A., Crusselle,V.J., Kang,S.-H.L., Engel,J.D. and Bungert,J. (2004) Recruitment of transcription complexes to the beta-globin gene locus in vivo and in vitro. *J. Biol. Chem.*, **279**, 50350–50357.

48. Vernimmen,D., Gobbi,M.D., Sloane-Stanley,J.A., Wood,W.G. and Higgs,D.R. (2007) Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J.*, **26**, 2041–2051.

49. Payankaulam,S., Li,L.M. and Arnosti,D.N. (2010) Transcriptional repression: conserved and evolved features. *Curr. Biol. CB*, **20**, R764-771.

50. Bergsland,M., Ramsköld,D., Zaouter,C., Klum,S., Sandberg,R. and Muhr,J. (2011) Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev.*, **25**, 2453–2464.

51. Vaxillaire,M., Rouard,M., Yamagata,K., Oda,N., Kaisaki,P.J., Boriraj,V.V., Chevre,J.C., Boccio,V., Cox,R.D., Lathrop,G.M., *et al.* (1997) Identification of nine novel mutations in the hepatocyte nuclear factor 1 alpha gene associated with maturity-onset diabetes of the young (MODY3). *Hum. Mol. Genet.*, **6**, 583–586.

52. Flanagan,S.E., De Franco,E., Lango Allen,H., Zerah,M., Abdul-Rasoul,M.M., Edge,J.A., Stewart,H., Alamiri,E., Hussain,K., Wallis,S., *et al.* (2014) Analysis of transcription factors key for mouse pancreatic development establishes NKX2-2 and MNX1 mutations as causes of neonatal diabetes in man. *Cell Metab.*, **19**, 146–154.

53. Xu,E.E., Krentz,N.A.J., Tan,S., Chow,S.Z., Tang,M., Nian,C. and Lynn,F.C. (2015) SOX4 cooperates with neurogenin 3 to regulate endocrine pancreas formation in mouse models. *Diabetologia*, **58**, 1013–1023.

54. Voight,B.F., Scott,L.J., Steinthorsdottir,V., Morris,A.P., Dina,C., Welch,R.P., Zeggini,E., Huth,C., Aulchenko,Y.S., Thorleifsson,G., *et al.* (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.*, **42**, 579–589.

55. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes (2008) *Genomics*, **91**, 243–248.

56. Ruf,S., Symmons,O., Uslu,V.V., Dolle,D., Hot,C., Ettwiller,L. and Spitz,F. (2011) Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat. Genet.*, **43**, 379–386.

747 57. Ding,X., Wang,X., Lin,M., Xing,Y., Ge,S., Jia,R., Zhang,H., Fan,X. and Li,J. (2016) PAUPAR lncRNA
748 suppresses tumourigenesis by H3K4 demethylation in uveal melanoma. *FEBS Lett.*, **590**, 1729–1738.

749 58. Akerman,I., Tu,Z., Beucher,A., Rolando,D.M.Y., Sauty-Colace,C., Benazra,M., Nakic,N., Yang,J.,
750 Wang,H., Pasquali,L., *et al.* (2017) Human Pancreatic β Cell lncRNAs Control Cell-Specific
751 Regulatory Networks. *Cell Metab.*, **25**, 400–411.

752 59. Engelkamp,D., Rashbass,P., Seawright,A. and van Heyningen,V. (1999) Role of Pax6 in development of
753 the cerebellar system. *Development*, **126**, 3585–3596.

754 60. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning
755 sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

756 61. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features.
757 *Bioinformatics*, **26**, 841–842.

758 62. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M.,
759 Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

760 63. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.,
761 Chou,A., Ienasescu,H., *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access
762 database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142-147.

763 64. Ward,L.D. and Kellis,M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and
764 regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930-
765 934.

766 65. Shen,Y., Yue,F., McCleary,D.F., Ye,Z., Edsall,L., Kuan,S., Wagner,U., Dixon,J., Lee,L., Lobanenkov,V.V.,
767 *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.

768 66. Halbritter,F., Vaidya,H.J. and Tomlinson,S.R. (2012) GeneProf: analysis of high-throughput sequencing
769 experiments. *Nat. Methods*, **9**, 7–8.

770 67. Hagège,H., Klous,P., Braem,C., Splinter,E., Dekker,J., Cathala,G., de Laat,W. and Forné,T. (2007)
771 Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat. Protoc.*, **2**,
772 1722–1733.

773 68. Gavrilov,A., Eivazova,E., Priozhkova,I., Lipinski,M., Razin,S. and Vassetzky,Y. (2009) Chromosome
774 conformation capture (from 3C to 5C) and its ChIP-based modification. *Methods Mol. Biol.*, **567**,
775 171–188.

776 69. Doyle,E.L., Booher,N.J., Standage,D.S., Voytas,D.F., Brendel,V.P., VanDyk,J.K. and Bogdanove,A.J. (2012)
777 TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target
778 prediction. *Nucleic Acids Res.*, **40**, W117–W122.

779

780 **Figure Legends**

**Figure1. H3K27ac profiling identifies regulatory elements in the PAX6 locus.** A. Diagram showing the human *PAX6* regulatory domain with flanking genes. Genomic position of known *PAX6* regulatory elements are marked in black (UCSC liftover from mouse coordinates) and sequence conservation (blue). B. Human *PAX6* locus with ChIP-seq for H3K27ac histone modification (green) in primary pancreatic islet and small intestine tissue from Human Epigenome Atlas, with MACS identified peaks (62). RNA-seq feature counts (black) for representative human β-cell sample (25) (Table S1), and vertebrate conservation (blue). C. ChIP-chip for H3K27ac histone modification in mouse β-TC3 cells (green), with called peaks. RNA-seq feature counts (black) for combined β-TC3 cell replicates. UCSC Refseq genes (orange) and known Pax6 regulatory elements. ChIP data represents average (n = 2). hg19 and mm9 coordinates.

**Figure 2. PE3 and PE4 cis regulatory elements drive neural and pancreatic expression in mouse embryos.** Stable mouse line with PE3 reporters showed strong signal in *Pax6* expressing tissues at E10.5 (A-B), E11.5 (C-D) while PE4 showed a more restricted *Pax6* expression pattern at E11.5 (E-F), E17.5 (G-H). A. Lateral view of E10.5 embryo showed LacZ staining in the ventral cerebral vesicles (cv) of the telencephalon, the ventral midbrain (mb), rhombencephalon (rc) and neural tube (nt). B. Transverse section through body of E10.5 embryo showed ventral (nt) staining in neurogenic region of the basal plate (bp), dorsal root ganglion (drg) and visible staining in the pancreas primoridum (pp). C. Lateral view of PE3 embryo at E11.5 showed ventral forebrain expression (vf), staining in (mb), the (rc) into the (nt) and the developing eye (e). D. Transverse section through body of E11.5 embryo revealed strong ventral (nt) staining in the neurogenic region of the basal plate (bp) of the (nt), plus the dorsal root ganglion (drg) surrounding the (nt), and (pp) in the body cavity. E. Lateral view of E11.5 embryo with staining in the forebrain (fb), ventral (mb), rhombic lip (rl) of rhombencephalon (rc), neural tube (nt) and (e). F. Transverse section through the body showed diffuse ventral neural tube (vnt) staining, dorsal root ganglion (drg), sympathetic chain neurons (scn) below the neural tube, and (pp) staining in the body cavity. G. Dissected E17.5 pancreas (p) cut into two cross sections revealed internal staining pattern. H. Lateral view dissected E17.5 brain showed nerve tracts in the forebrain consisting of lateral olfactory tracts (lot) and ventral brain nuclei staining in cerebellum (cb) and hindbrain (hb).

**Figure 3. Histone modification and transcription factor binding across PE3 at the *Pax6* locus in mESC and characterization of a PE4 regulatory variant** A. 100 kb genomic region surrounding the PE3 element at the *Pax6* gene showing histone modifications and transcription factor binding in mESC (32). B. Position of putative regulatory SNP in Pax4 binding sequence, with motif and alignment score from Jaspar database (63). C. (left) Western blot analysis of β-TC3 cells transfected with FLAG-PAX4 or FLAG-hnRNPU with a mock transfection control, detected with an antibody against FLAG and using histone H3 antibody as a loading control. (right) Western blot analysis of β-TC3 cells transfected with untagged PAX6 or PAX4 cDNAs and a mock transfection control, probed with antibodies against PAX4, PAX6 and hnRNPU as a loading control. D. ChIP assay for FLAG-PAX4 in β-TC3 cells, evaluated by qPCR, using primer sets to the mouse PE4 element. Top, Schematic showing primer locations, with negative control primers to a sequence 4 kb upstream of PE4 (PE4_Con), intergenic regions on Chr2 (Chr2_con) and a large intron in Pola1 on HSAX (chrX_Con). Values are displayed as percentage of input sample, with two independent experimental replicates of FLAG-PAX4 ChIP. Error bars +/- SEM. E. Dual luciferase assay for human PE4 element encoding the G or C variant in pGL4.23 vector, transfected into β-TC3 cells, co-transfected with human PAX4 cDNA. Relative luciferase signal represents firefly luciferase over Renilla luciferase signal, normalised to the signal from the empty vector. Error bars +/-SEM, n=5; P value from Welch's two sample t-test (p values: **** < 0.0001, *** < 0.001, ** < 0.01, * < 0.05).

**Figure 4. 3C-qPCR reveals PE3 and PE4 interaction with multiple regions over *Pax6* gene.** A. PE3 3C-qPCR data displayed as relative cross-linking frequencies between the PE3 anchor fragment and Bgl*II* fragment primers across the mouse *Pax6* locus in β-TC3 and mESC. B. PE4 3C-qPCR data from the PE4 anchor fragments in β-TC3 and mESCs. Data sets are scaled to one another and display genomic position relative to 3C primer values. Data from the two cell lines were normalised using a probe located in the *Ercc3* locus. Signal is the combined mean relative cross-linking frequency of 6-8 individual qPCR values, across 2 experimental replicates. Error bars +/-SEM.

**Figure 5. TALE recruitment to regulatory elements can modulate target gene expression.** A. Schematic showing TAL effector proteins designed to PE3 and PE4 regulatory elements fused to activator VP64 or repressor SID4X domains, with co-expressed GFP. Position of PE3 and PE4 TAL effectors with respect to CTCF and PAX4 binding sites. B-D. Graph showing relative change in *Pax6*, *Paupar*, *Elp4, Immpl1 and Rcn1* and *Oct4* expression after transfection with TAL effectors fused to VP64 or SID4X targeted to the PE3

or PE4 regulatory elements in β-TC3 or mESC. RNA levels were quantified by RT-PCR and normalised to 18S RNA. Error bars +/-SEM, n=2. P value from a Welch t-test of ΔCT values (p values: **** < 0.0001, *** < 0.001, ** < 0.01, * < 0.05).

## Abbreviations

(BMI) Body Mass Index

(3C) chromosome conformation capture

(CNS) central nervous system

(DRR) downstream regulatory region

(ESC) mouse embryonic stem cells

(lncRNAs) long noncoding RNAs

(P/EE) Pax6 pancreas and ectoderm enhancer

(PE2) Pax6 pancreases cis-regulatory element 2

(PE3) Pax6 pancreases cis-regulatory element 3

(PE4) Pax6 pancreases cis-regulatory element 4

(SNP) Single Nucleotide Polymorphisms

(TAD) Topologically associated domain

(TAL) Transcription activator-like

(T1D) Type 1 diabetes

(T2D) Type 2 diabetes